National Aeronautics and Space Administration

Schedule and Abstracts

Conference on Intelligent Data Understanding

AISR

NASA Applied Information Systems Research Program
Principal Investigators Meeting

October 14-16, 2009

NASA Ames Research Center
Moffett Field, CA

Sponsored by

NASA Aviation Safety Program
NASA Advanced Information Systems Research Program

www.nasa.gov

Conference on Intelligent Data Understanding

AISR
NASA Applied Information Systems Research Program
Principal Investigators Meeting

October 14-16, 2009

NASA Ames Research Center
Moffett Field, CA

This year NASA's Conference on Intelligent Data Understanding (CIDU) is being held in conjunction with the Applied Information Systems Research Program Principal Investigators meeting. This conference brings together top researchers for talks in the following areas:

- Autonomous Systems and Systems Health Management
- Discovery Algorithms
- Data/Information Infrastructure
- Computational Methods and Frameworks

Organizing Committee

Ashok Srivastava, Ph.D.	NASA Ames Research Center
Nikunj Oza, Ph.D.	NASA Ames Research Center
Joe Bredekamp	NASA Headquarters
Jeff Scargle, Ph.D.	NASA Ames Research Center
Ramasubbu Venkatesh, Ph.D	Netflix Inc.

Conference Agenda

Wednesday, October 14

7:30 AM	Registration/Continental Breakfast
8:00 AM	*Opening Remarks –* Joe Bredekamp, NASA Headquarters (AISRP) Ashok Srivastava, NASA Ames Research Center (CIDU)
8:30 AM	Autonomous Systems and Systems Health Management
	Introduction to Autonomous Systems
CIDU	*Creating Knowledge from IT System Events* – Ira Cohen, Hewlett Packard
CIDU	*Fingerprinting the Datacenter: Online Classification of Performance Problems* – Moises Goldszmidt, Microsoft Research
AISRP	*An Analytical Tool for Robot Mission Reliability Prediction* – John M. Dolan, Carnegie Mellon University
10:00 AM	* ***break*** *
10:30 AM	Discovery Algorithms
	Introduction to Discovery Algorithms
CIDU	*PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce* – Sugato Basu, Google
AISRP	*Stochastic Models for High-Dimensional, Nonstandard Data* – Chad Schafer, Carnegie Mellon University
AISRP	*Automatic Detection of Sub-Kilometer Craters in High Resolution Planetary Images* – Wei Ding, University of Massachusetts
AISRP	*A Novel Higher-Order Statistical Method for Extracting Dependencies in Multivariate Geospace Data Sets* – Simon Wing, Johns Hopkins University
12:00 PM	* ***Lunch*** *
1:00 PM	**Keynote Address**
AISRP	*The Challenge of Digital Astronomy* – Roger Blandford, Kavli Institute for Particle Astrophysics and Cosmology, Stanford University
2:00 PM	Discovery Algorithms (cont.)
AISRP	*Novel Methods for the Analysis of Photon-Limited Data* – Jeff Scargle, NASA Ames Research Center
AISRP	*Object-Based Image Analysis Utilizing Image Segmentation Hierarchies* – James C. Tilton, NASA Goddard Space Flight Center
AISRP	*Streaming the Sky: High Performance Knowledge Discovery* – Andrew Connolly, University of Washington
3:00 PM	* ***break*** *
3:30 PM	Discovery Algorithms (cont.)
	Introduction
AISRP	*Automated, Dynamical Event Classification and Response in a Robotic Sensor Network* – S. G. Djorgovski, Caltech
AISRP	*GLYDER: Global Detection of Cyclones from Multi-Satellite Data* – Ashit Talukder, NASA Jet Propulsion Laboratory

AISRP	*Scalable Algorithms for Fast Analysis of Megapixel CMB Maps and Large Astronomical Databases* – Istvan Szapudi, University of Hawaii
AISRP	*Detecting Transient Surface Features with Dynamic Landmarking* – Kiri L. Wagstaff, NASA Jet Propulsion Laboratory
AISRP	*Enabling Bayesian Inference for the Astronomy Masses* – Martin Weinberg, University of Massachusetts
5:00	End of Day One
6:00 PM	**ACM Presentation** at the Exploration Center

Thursday, October 15

7:30 AM	Registration/Continental Breakfast
8:00 AM	*Opening Remarks* – Joe Bredekamp
8:15 AM	Discovery Algorithms (cont.)
	Introduction to Discovery Algorithms Wrap-up
AISRP	*Algorithmic Integration of High-throughput Data and Metabolic Models* – Stephen Fong, Virginia Commonwealth University
AISRP	*Automated Orbital Mapping* – David Wettergreen, Carnegie Mellon University
AISRP	*Discovery with MTool on Massive High Dimensional Data* – Martin Lo, NASA Jet Propulsion Laboratory
AISRP	*Distributed and Peer-to-Peer Data Mining for Scalable Analysis of Data from Virtual Observatories* – Hillol Kargupta, University of Maryland, Baltimore County
CIDU	*From Learning to Knowledge Discovery to Action in Distribution Sensitive Scenarios* – Nitesh Chawla, University of Notre Dame
10:00 AM	* ***break*** *
10:15	Data/Information Infrastructure
	Introduction to Data/Info Infrastructure
CIDU	*Dedicated Decision Systems* – Eric Colson, Netflix
AISRP	*NASA's Cosmos* – Kenneth R. Lang, Tufts University
AISRP	*Magnetogram Synthesis: A Vital Data Analysis Component of A Space Weather Prediction Infrastructure* – Joel Allred, Drexel University
AISRP	*Advanced Visualization in Solar System Exploration and Research (ADVISER): Optimizing Science Return from the Moon and Mars* – John Huffman, Brown University
AISRP	*Automated DTM Generation for HiRISE and LROC* – Zachary Moratto, Stinger Ghaffarian Technologies
12:00 PM	* ***Lunch*** *
1:30 PM	**Poster Session**
3:00 PM	* ***break*** *
3:30 PM	Data/Information Infrastructure (cont.)
	Introduction
CIDU	*The Cloud Goes BOOM: Data-Centric Programming for Datacenters* – Joe Hellerstein, University of California, Berkeley

AISRP	*Visualization of Terascale Datasets with Impostors* – Thomas Quinn, University of Washington
AISRP	*Large Scale On Demand Cross-Matching with Open SkyQuery* – Ani R. Thakar, Johns Hopkins University
AISRP	*On-the-fly and Grid Analysis of Astronomical Images for the Virtual Observatory* – Andrew Ptak, Johns Hopkins University
AISRP	*Martian Tactical Geologic Mapping for Mars Surface Missions* – Mark Powell, NASA Jet Propulsion Laboratory
5:00	End of Day Two

Friday, October 16

7:30 AM	Registration/Continental Breakfast
8:00 AM	*Opening Remarks* – Ashok Srivasatava
8:15 AM	**Keynote Address**
CIDU	*Social Technology* – Marti Hearst, University of California, Berkeley
9:15 AM	*Computational Methods and Frameworks*
	Introduction to Computational Methods
AISRP	*Integration of Orbital, Descent and Ground Imagery for Topographic Capability Analysis in Mars Landed Missions* – Rongxing (Ron) Li, Ohio State University
AISRP	*GPU-based Tools for Computational Astrophysics: N-Body Integrators for Dynamical Systems* – Mario Juric, Institute for Advanced Studies
AISRP	*Controlling Sensitive Trajectories, Mission Extension, and Material Transfer Between Planetary Systems* – Edward Belbruno, Princeton University
10:15 AM	* ***break*** *
CIDU	*Hadoop: Distributed Data Processing* – Amr Awadallah, Cloudera
AISRP	*Development of an Adaptive Non-Ideal MHD Simulation Tool for Multiple Space Science Applications* – Gabor Toth, University of Michigan
AISRP	*Parallel-Processing Astrophysical Image-Analysis Tools* – Kenneth Mighell, National Optical Astronomy Observatory
11:30 AM	* ***Lunch*** *
12:30 PM	*Computational Methods and Frameworks (cont.)*
AISRP	*EPISODE: Software for Trajectory Generation* – Jeff Jewell, NASA Jet Propulsion Laboratory
AISRP	*How Well Do You Know That? Uncertainty Analysis in Earth Remote Sensing* – Robin Morris, RIACS
AISRP	*Astronomy in the Cloud* – Jeffrey P. Gardner, University of Washington
AISRP	*Directed Exploration of Complex Systems* – Michael Burl, NASA Jet Propulsion Laboratory
AISRP	*Demonstration of Safe Human/Robot Coordination on the Athlete Lunar Rover* – Brian C. Williams, Massachusetts Institute of Technology
1:45 PM	*Concluding Remarks*

Presentations

Keynote Addresses

Autonomous Systems and Robotics

Discovery Algorithms

Data/Information Infrastructure

Computational Methods and Frameworks

Invited Presentations

Keynote Address

The Challenge of Digital Astronomy

Roger Blandford
Kavli Institute for Particle Astrophysics and Cosmology, Stanford University

The practice of astronomy is being revolutionized by the exploitation of modern computational methods. Theoretical astrophysics is increasingly concerned with numerical solutions to problems that are beyond the scope of conventional analysis and large-scale simulations that often contain detailed observational programs. Modern telescopes, both those in operation and those that have been proposed, generate data at prodigious rates and pose an organizational challenge that may even surpass the demands that they are making on hardware in a "green" future. New paradigms for planning facilities archiving the data and mining databases are emerging with implications for the astronomical workforce of tomorrow.

Keynote Address

Social Technology

Marti Hearst
University of California, Berkeley

We are in the midst of extraordinary change in how people interact with one another and with information. A combination of advances in technology and change in people's expectations is altering the way products are sold, scientific problems are solved, software is written, elections are conducted, and government is run.

People are social animals, and as Shirky notes, we now have tools that are flexible enough to match our in-built social capabilities. Things can get done that weren't possible before because the right expertise, the missing information, or a large enough group of people can now be gathered together at low cost.

These developments open a number of interesting research questions and potentially change how scientific research should be conducted. In this talk I will attempt to summarize and put some structure around some of these developments.

Creating Knowledge from IT System Events

Ira Cohen
Hewlett Packard

IT systems produce large quantities of events, describing the state of various components. These events can be used to detect and troubleshoot various problems, such as performance and security. They can also be useful in designing improvements to a system, and in gaining general understanding of system behavior.

In my talk I will describe the challenges in leveraging these events in today's environments, and describe algorithmic and technological solutions we developed for overcoming these challenges.

Fingerprinting the Datacenter: Online Classification of Performance Problems

Moises Goldszmidt
Microsoft Research

Due to scale, utilization, and other factors, large distributed systems in data centers such as the ones supporting web services and cloud computing may experience recurrent performance problems. In order to maintain the desired quality of service and avoid interruptions, automatic recognition of the problems for rapid diagnosis and resolution is a must. This is non-trivial task, given the large amount of data in the monitoring of these systems. In this talk I will present recent research on methods and mechanisms to address this challenge, and validate this research with results on real data from one of Microsoft's online services.

An Analytical Tool for Robot Mission Reliability Prediction

John M. Dolan
Carnegie Mellon University

Space missions require robots that are highly reliable for the smallest possible cost. Current theory and practice do not provide a methodical approach to choosing a set of robot modules with demonstrably high reliability at minimal cost. In particular, there is no methodology clearly linking mission specifications, to include desired reliability, to an appropriate choice of robot modules for composing a robot or multi-robot team. The objective of our work is therefore to establish a principled analytical methodology for the inclusion of reliability considerations in mobile robot and multiple mobile robot team mission design. This work is a component of the longer-term project of creating comprehensive, principled analysis tools for general mobile robot design, to include mobility, weight, speed, kinematic, and other characteristics.

This talk focuses on the creation of a general taxonomy for mission specification, cost-reliability tradeoffs for mission designs in various parts of the mission taxonomy space, and the application of reliability considerations to task allocation in multi-robot missions. We show that the inclusion of more realistic reliability considerations (i.e., the notion that a robot has a non-zero likelihood of failure in completing its task) changes the optimal task allocation in a large percentage of cases, even when using an incomplete heuristic mission planner.

PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce

Sugato Basu
Google

Classification and regression tree learning on massive datasets is a common data mining and discovery task, yet many state-of-the-art tree learning algorithms require training data to reside in memory on a single machine. While more scalable implementations of tree learning have been proposed, they typically require specialized parallel computing architectures. In contrast, the majority of Google's computing infrastructure is based on commodity hardware. In this talk, we describe PLANET, a scalable distributed framework for learning tree models over large datasets. PLANET defines tree learning as a series of distributed computations, and implements each one using the MapReduce model of distributed computation. We show how this framework supports scalable construction of classification and regression trees, as well as ensembles of such models. We discuss the benefits and challenges of using a MapReduce compute cluster for tree learning, and demonstrate the scalability of this approach by applying it to a real-world learning task.

Stochastic Models for High-Dimensional, Nonstandard Data

Chad Schafer
Carnegie Mellon University

We describe nonparametric techniques for density estimation for high-dimensional data which exhibit irregular dependence structure that cannot be modeled by parametric multivariate distributions. In such situations, a low-dimensional data representation is critical because of the curse of dimensionality. Our proposed methodology depends on dimensionality reduction and data parameterization, in which the data of nonstandard form are mapped into a space where standard techniques can be used for density estimation and simulation. The algorithm is illustrated via an application to tropical cyclone spatial variability in the North Atlantic; each datum in this case is an entire hurricane trajectory.

This is joint work with Susan Buchman, Peter Freeman, and Ann Lee.

AUTOMATIC DETECTION OF SUB-KILOMETER CRATERS IN HIGH RESOLUTION PLANETARY IMAGES

Wei Ding
University of Massachusetts

Impact craters are the most studied geomorphic features in the solar system because they yield information about the past and present geological processes on planetary surfaces. We have developed and implemented a crater detection algorithm capable of effective detection of sub-kilometer craters from high-resolution panchromatic images. The algorithm is based on combination of mathematical morphology, Haar-like texture descriptors, and supervised machine learning. The algorithm has been tested on a large High Resolution Stereo Camera (HRSC) 12.5-m/pixel image centered on the Nanedi Valles, Mars. The concept of using mathematical morphology/machine learning combination for crater identification has been validated. Dependences of detection rate, branching factor, and quality rate on minimum size of detected craters, as well as the receiver operating characteristic curve (ROC) for our algorithm has been established. Our crater detection algorithm achieves up to 80% detection percentage for craters with diameter > 200 m. A pipeline for cataloging craters in a large image, which needs to be tessellated into a number of tiles, has been developed. This pipeline was applied to the entire HRSC image, resulting in a catalog of over 35,000 sub-kilometer craters.

A Novel Higher-Order Statistical Method for Extracting Dependencies in Multivariate Geospace Data Sets

Simon Wing
Johns Hopkins University

We developed a number of tools based on the discriminating statistics of entropy-based measures of information that have proven useful for analysis of geospace data. Using mutual information and cumulants, we were able to detect nonlinearity in the response of the Kp index driven by solar wind input. The nonlinear response occurred on a timescale around 50 hours and appears to be associated with high-speed solar wind streams. To examine causality in space plasmas, we used the transfer entropy as a measure of causal dependence. We found that transfer entropy, unlike cross-correlation, is one-directional and can be used to determine the extent to which variables are coupled even if they are highly correlated because of a common driver. We have also explored how cumulant-based measures can be used to improve predictive modeling. Cumulant-based significance can be used as a discriminating statistic to identify the presence of noise in a system. Noisy data can be eliminated from the training dataset, so as to improve model predictions. We applied this technique to a nonlinear time series to which we added various levels of noise. Although a neural network trained on the noisy dataset was not accurate, when noisy data was eliminated using the cumulant-based significance, the output was nearly as good as a neural network trained on a clean data set. We also examined which discriminating statistics give the best results when a dataset is contaminated by noise or has a limited number of measurements. Finally, we developed a nonlinear predictive model for the Kp and Dst geomagnetic indices, which measures the strength of geomagnetic activity. We found that our predictive model also exhibits a solar cycle dependence in accuracy, which we attribute to the presence of stronger nonlinearity during the solar cycle minimum. Our Kp and Dst models have been adapted by national space weather centers.

Novel Methods for the Analysis of Photon-Limited Data

Jeff Scargle
NASA Ames Research Center

Statistical techniques, and efficient algorithms to implement them, will be developed for analysis of event data (a.k.a. point data) in astrophysics, space and Earth science, and other applications. These methods, integrated with existing standard ones, will be described in a Handbook of Statistics for Event Data, providing scientists with practical guidance on use and interpretation. This choice of problem area is based on my experience and expertise, and deliberately focuses on an important class of NASA missions. Event data are largely photon detections: photometry, imaging, timing, and spectrophotometry. Most important in high-energy astrophysics, where arrival direction and time, plus the energy of individual photons can be accurately measured, the methods will also have broader applicability. The specific problems addressed include repeated measurements; detection, characterization, and upper limits (for signals consisting of spatial sources, time variability, or spectral lines); and analysis of image and other structure. Algorithm design will focus on the practical issues of user comprehension of inputs and processing, plus understanding and interpretation of the outputs. Statistically rigorous modeling will be combined with robust applicability to diverse contexts, ease of use, transparency of parameter adjustment, validity and applicability of approximations, unification of novel and old methods, and computation efficiency. Matlab algorithms (with adaptations to other high-level systems in common use, such as IDL, R, S, and Python, as well as low-level compiled languages) will posted in the AISRP code repository, with documentation there and in the accompanying Handbook.

Object-Based Image Analysis Utilizing Image Segmentation Hierarchies

James C. Tilton
NASA Goddard Space Flight Center

Currently available pixel-based analysis techniques do not effectively extract the information content from the increasingly available high spatial resolution remotely sensed imagery data. A general consensus is that object-based image analysis (OBIA) is required to effectively analyze this type of data. OBIA is usually a two-stage process; image segmentation followed by an analysis of the segmented objects. We are exploring approaches to OBIA based on the analysis of hierarchical image segmentations provided by the Recursive Hierarchical Segmentation (RHSEG) software developed at NASA GSFC. We discuss a general OBIA approach in which RHSEG-produced image segmentation hierarchies are analyzed by the Subdue graph-based knowledge-discovery system developed by a team at Washington State University. We describe our initial approach to representing the RHSEG-produced hierarchical image segmentations in a graphical form understandable by Subdue, and provide results on real and simulated data. We also discuss planned improvements designed to more effectively and completely convey the hierarchical segmentation information to Subdue and to improve processing efficiency.

Streaming the Sky: High Performance Knowledge Discovery

Andrew Connolly
University of Washington

With the development of large multispectral surveys and initiatives such as the Virtual Observatory (VO), the last decade has witnessed a change in the way astronomers work with data. The federation of these massive data streams provides an opportunity for astronomers to study the origin, structure, and evolution of the universe with unprecedented accuracy. It does, however, come the challenge of how to enable access to these data in a way that will enhance the scientific return. We describe here a new way to interface services and data streams; one that builds upon emerging Web 2.0 technologies for streaming large amounts of data with personalization of access to data through the use of gadgets and widgets embedded within browser-configurable clients.

Automated, Dynamical Event Classification and Response in a Robotic Sensor Network

S. G. Djorgovski
Caltech

A new generation of scientific measurement systems is emerging in astronomy, space science, and many other fields: connected sensor networks which gather and analyze data automatically, and respond to outcome of these measurements in real time, often redirecting the measurement process itself, and without human intervention. Examples from the rapidly developing field of time domain astronomy include discovery of moving objects (e.g., potentially hazardous, Earth-crossing asteroids) or transient or explosive astrophysical phenomena (supernovae, gamma-ray bursts, etc.), each requiring rapid follow-up observations. Their detection and characterization leading to an automated follow-up decision and additional measurements must be done in a fully automated and reliable fashion. We describe some novel approaches to automated astrophysical event classification in digital synoptic sky surveys.

GLYDER: Global Detection of Cyclones from Multi-Satellite Data

Ashit Talukder
NASA Jet Propulsion Laboratory

Tropical and extra-tropical cyclones are key manifestations of the oceanic air-sea interaction and contribute to regional heat exchanges, which affects ocean and atmosphere dynamics. To identify and track tropical weather systems, the Tropical Prediction Center/National Hurricane Center (TPC/NHC) uses conventional surface and upper-air observations and reconnaissance aircraft reports from field missions. These observations, however, are concentrated in the North American coasts and in Japan/Europe to some degree. Coverage on a global basis, especially in under-developed and developing nations such as large portions of Asia and Africa, is limited or lacking. In recent years, some studies have used satellite images that are manually retrieved and analyzed to improve the accuracy of cyclone tracking. This procedure is currently slow, tedious, and requires close analysis by teams of experts. Incorporating multiple sensor types together provides several cues that can better characterize cyclones. In GLYDER, we address this shortcoming by using new data processing and machine learning methods on multiple remote satellite data that enables detection and tracking of cyclones from multisensor data and visualize these tracks at the finest temporal and spatial resolutions; such cyclone tracking capabilities by transferring knowledge across multiple sensors have not been demonstrated in the past.

Our latest advances include several fundamental theoretical novelties and integration of software tools to build a usable system for global cyclone tracking. We have designed a new multisource machine learning theory that enables transfer learning and knowledge sharing between multiple disparate data streams. This allows us to detect cyclones more effectively, and maintain tracks of their movement with an unprecedented temporal accuracy of ½ hour. We have also designed algorithms to robustly identify cyclones from three sources, QuikSCAT, TRMM, and GOES. We have initiated algorithmic modifications to handle processing of near-real-time data (that are noisier and have missing measurements) to enable tracking of real-time cyclonic events. More recently, we have designed new data mashup tools and statistical tools to include RSS feeds from different sources, to improve the accuracy of our GLYDER cyclone tracking tool for real-time data streams. We have developed software tools to automatically retrieve satellite data from different sources, preprocess them (data gridding), and extract cyclone location information from such data. New visualization tools enable end users to type specific queries (such as searching for cyclones over range of dates and in particular regions) and visualize the tracks of multiple cyclones as they evolve in space and time. We have engaged in discussions with several groups to determine the best path forward for dissemination of the GLYDER cyclone tracking toolkit to the science community.

In the near future, we plan to conduct further tests of GLYDER on larger historical datasets and near-real-time streams for validation. All algorithmic components will be improved to ensure more robust performance and capabilities will be added to handle missing measurements and higher noise in real-time streams. We will also plan to explore collaborations with field campaign teams and potentially release the GLYDER software through appropriate channels by the end of the project.

Scalable Algorithms for Fast Analysis of Megapixel CMB Maps and Large Astronomical Databases

Istvan Szapudi
University of Hawaii

As computing power scales only linearly with the (exponentially growing) data rate, higher-order statistics, which scales as some power k of the data items, has been a daunting challenge. I will review approximate techniques for calculating higher-order correlation functions and related quantities that scale close to linear. These algorithms enable the practical calculation of these statistics with reasonable computing resources even with large (megapixel+) datasets.

Detecting Transient Surface Features with Dynamic Landmarking

Kiri L. Wagstaff
NASA Jet Propulsion Laboratory

Our goal is to autonomously detect transient surface features, such as dust devil tracks or dark slope streaks on Mars, from images. In contrast to pixel-based change detection, our approach operates on landmarks, which are visually salient regions within an image.

Landmarks provide a sparse, semantic representation of the image content. We have developed automated methods to identify salient landmarks in each image (using statistical salience measures), characterize and classify them (using a machine learning classifier), and then match landmarks between images (using a Relative Landmark Graph and graph-matching algorithms). New or vanished landmarks are detected when they fail to match between images. Matched landmarks can be compared to detect any changes in the landmark attributes. We have demonstrated the ability to identify landmarks and detect changes on images collected by several different Mars cameras, such as the Mars Orbiter Camera (MOC) on Mars Global Surveyor and the Thermal Emission Imaging System (THEMIS) on Mars Odyssey.

Enabling Bayesian Inference for the Astronomy Masses

Martin Weinberg
University of Massachusetts

I will present a short introduction to the UMass Bayesian Inference Engine (BIE) followed by progress in galaxy properties inference semi-analytic models and star count/population analysis. I will describe two of these at length. GALPHAT: our BIE-based package written to perform two-dimensional image. Our combination of accurate optimized numerics with full posterior distributions for parametric image decomposition allow inferences using Bayes Factors over a wide variety of competing models and hypotheses. Recent progress includes a new Bayes Factor computational technique. BIE-SAM: semi-analytic models (SAMs) attempt to learn the importance of the modeled physical mechanisms in shaping galaxies by testing the degree of agreement between the model prediction and the observation. To incorporate SAM into the framework of Bayesian inference, we have built a SAM, incorporating most of the mainstream prescriptions so that we can explore not only in a big parameter space but also various proposed parameterizations. The aim of our approach of introducing the Bayesian inference into semi-analytic is to put SAM investigation on a rigorous probabilistic footing.

Algorithmic Integration of High-throughput Data and Metabolic Models

Stephen Fong
Virginia Commonwealth University

The concurrent development of high-throughput experimental data and large-scale computational models potentially allows for organisms to be studied coherently as intact systems. Challenges still remain on how best to use the existing data and computational techniques. Due to the large quantity of information generated by high-throughput analyses, it is often difficult to surmise the functional consequences of measured changes. Computationally, it is often possible to predict functional consequences, but the accuracy to in vivo reality is sometimes in question. In this talk, we will discuss a framework for integrating experimental high-throughput data with metabolic models in an attempt to produce results associated with function that are reflective of a true biological state. Examples of applying this framework include using proteomic data to predict stage-specific life cycle functions in parasites and transcriptomic data to predict differences between cells with different phenotypes (normal vs. abnormal stem cells).

Automated Orbital Mapping

David Wettergreen
Carnegie Mellon University

Planetary exploration has entered an era in which our data-gathering ability has outpaced our capacity for timely analysis. A combination of high-resolution instruments, spacecraft autonomy and mobility, and extended lifetimes is adding to a vast and growing library of image and spectral data.

In this research we address the bottleneck in data analysis by creating methods to automatically generate maps from hyperspectral data. Our method exploits recent advances in machine learning and object recognition. The automated mapping system is learning to infer mapping rules statistically from "training" data in the form of existing geologic maps and registered orbital images. This is a natural application of existing learning-based methods of image segmentation and boundary detection. We will extend these methods to the orbital mapping domain with geologically relevant descriptors like texture, spectral signatures, and surface features like boulders, craters, and dunes. The result is a tool that uses features to predict the locations of geologic units and their boundaries.

Automating geologic mapping will yield immediate benefits for the science community. It will dramatically improve the speed of image analysis which is limited by manual feature labeling. Maps offer fast summaries of gigapixel images that may include hundreds of thousands of features. These "draft" maps can be refined by human experts in less time than it would take to construct an entire map from scratch.

Automated mapmaking is not intended to replace human expertise; both human and automated image analysis play to different strengths. The human can offer expert geologic insight and authoritative segmentations of geologic units. The information system can examine many images quickly, quantify features with accuracy and consistency, and perform exhaustive statistical analysis to search for unexpected patterns and anomalies. In this sense the roles of the geologist and the automated mapping system are strongly complementary. The mapping system automates some of geologists' more arduous tasks, freeing them to focus on interpretation.

Discovery with MTool on Massive High Dimensional Data

Martin Lo
NASA Jet Propulsion Laboratory

The goal of MTool is to provide researchers and engineers with a suite of tools to analyze the geometric and topological structures hidden within their data. More often than not, these coherent structures are dynamically changing in time. Moreover, the structures typically live in high-dimensional phase spaces with complicated coordinates which are not readily computable or visualizable. This is where tools in computational differential geometry and algebraic topology can be used to extract and analyze these hidden structures. We are currently working on four different applications: tomographic imaging of astrophysical ßgamma ray sources using Radon transforms, simulations of bar galaxy structures using invariant manifolds, analysis of large-scale structures of the universe using homology theory, and the extraction of Lagrangian coherent structures in climate data using dynamical systems theory.

Distributed and Peer-to-Peer Data Mining for Scalable Analysis of Data from Virtual Observatories

Hillol Kargupta
University of Maryland, Baltimore County

Design, implementation, and archival of very large sky surveys are playing increasingly important roles in today's astronomy research. Current projects such as GALEX All-Sky Survey and future ones such as WISE All-Sky Survey are destined to produce enormous catalogs of astronomical sources. The Large Synoptic Survey Telescope is supposed to stream in large volumes of data at a high rate. These distributed data sources, a community of users distributed all over the world, and systems such as MyDB, MySpace in AstroGrid, and Grid Bricks for storing and managing users' local data are opening up the possibility of constructing Peer-to-Peer (P2P) networks for data sharing and mining. This research is exploring the possibility of using distributed and P2P data mining technology for exploratory astronomy from data integrated and cross-correlated across these multiple sky surveys. The research has so far developed P2P collaborative classifier learning algorithms, distributed outlier learning algorithm, and distributed eigenstate monitoring algorithms. The algorithms are used for various astronomy numeric and text data mining problems in a P2P environment. This talk will offer an overview of the research accomplishments.

From Learning to Knowledge Discovery to Action in Distribution Sensitive Scenarios

Nitesh Chawla
University of Notre Dame

Models for knowledge discovery in the real world face the pervasive and compelling problem of irregularities in data distribution. Decisions that are optimal in expected utility can be vulnerable to failure, and value functions that reflect the discontinuities of the real-world pragmatics can quickly become intractable. Surprises can happen in uncertain environments. The class distributions may not be the same (imbalanced data), with the class of interest being rare or extreme. The training and testing distributions can differ. The costs of making mistakes or benefits from making correct predictions may also not be constant and can evolve due to operational reasons. I will present some of our work on tackling such challenges in the journey from data to learning to knowledge discovery to action.

Dedicated Decision Systems

Eric Colson
Netflix

Transaction processing systems are used in the daily operations of most modern businesses. They have evolved to execute flawlessly at unbelievable scale. However, perfect execution does not necessarily equate to optimal or even efficient operations. Achieving strategic alignment with business objectives requires more than execution. Operations needs a system that is capable of applying judgment, picking the best alternative available in order to achieve stated goals. A system for such operational decision-making warrants a fundamentally different approach from that of the passive transaction processing systems. This talk will discuss how to define and position a decisions system as a deliberate and dedicated component of operations. The talk with draw upon experiences from Netflix's DVD operations.

NASA's Cosmos

Kenneth R. Lang
Tufts University

This presentation will discuss NASA's Cosmos, a website that provides comprehensive accounts of the most recent discoveries about the planets, their satellites, the Sun, and other bodies in the solar system, based primarily on NASA space missions. NASA's Cosmos includes printed books, E-books, and lectures, and has developed a web site, at http://ase.tufts.edu/cosmos/, that presents NASA scientific results to the general public, space scientists, students, and teachers. The books convey the scientific credibility, historical authority, human interest, and visual excitement of NASA Space Science. Numerous experts in the field have read all the draft documents for accuracy and completeness, and they are accessible to all readers, from interested laypeople to students or professionals. Some of these books have been translated into French, German, Italian, and Japanese, and the most recent E-books include electronic access to thousands of seminal reference papers.

Magnetogram Synthesis: A Vital Data Analysis Component of A Space Weather Prediction Infrastructure

Joel Allred
Drexel University

Space weather is the term used to describe the numerous phenomena that result from highly energetic solar particles impacting on the Earth's atmosphere. Transient magnetic events in the solar atmosphere are ultimately the source of all space weather. Thus, models which attempt to predict space weather must have an accurate global picture of the current state of the Sun's magnetic field. In order to provide a more complete picture of the Sun's magnetic topology, we are creating a tool which can stitch together observations from a disjointed and unevenly sampled data supply using all available sources. This tool can manage the acquisition of data from the large number of observatories monitoring the solar magnetic field, cast them into a source-independent format, and then perform the spatial and temporal interpolation necessary to provide the magnetic field values needed by models. We have called this tool MAGIC (MAGnetogram Integration and Composition). In order for MAGIC to be useful for science and forecasting analysis, it must be capable of doing a host of post-processing computations. MAGIC supports two post-processing layers. The first is a lightweight layer which provides a number of relatively simple analyses such as monopole subtraction, interpolation, and line-of-sight to radial magnetic field projection. A second heavyweight layer provides an interface to third-party processing tools. For example, MAGIC will provide an interface to tools which can estimate surface flow velocities from the evolution of the magnetic field and can perform coronal field extrapolation from photospheric magnetograms. Here we report on the progress in developing the MAGIC tool.

Advanced Visualization in Solar System Exploration and Research: Optimizing Science Return from the Moon and Mars

John Huffman
Brown University

ADVISER (Advanced Visualization in Solar System Exploration and Research) draws from our previous successes creating compelling visualization tools for advanced geosciences research using fully immersive virtual reality systems. The overall goal of ADVISER is to virtually place the geoscientist on planetary surfaces, and provide the tools necessary to explore and take measurements within this environment. This has proven to be useful enough in exploring planetary surfaces that it has now been incorporated into the Brown University Geology course "Mars, Moon, and Earth."

A significant effort has been placed on adapting the tools to explore planetary datasets tens of terabytes in size, while maintaining interactive rendering speeds needed for virtual reality. The tools are being developed both for our advanced immersive visualization systems and as commodity based systems. Potentially, this allows for a larger user base, without compromising usability or functionality. This also allows for seamless transition from simple desktop visualization, group visualization at high-resolution display walls, and fully immersive visualization in the Cave. On the desktop, the tools are invoked from within ArcMap, allowing direct access to ADVISER from within a geologists known toolset.

A key aspect of the ADVISER project is the idea of Data Fusion: the ability to combine multiple forms of data into a single unified form. While this includes combining geo-referenced height and image data into a seamless landscape, it also includes incorporating the vast amounts of multi-spectral data being collected. Simulations also allow the researcher to explore active geological processes, as well as the effects of solar insolation. Data are combined into a single, coherent visualization geoscientists can interact with as though they were on or near the surface of remote locations like Mars or Antarctica. ADVISER provides tools such as strike and dip, real-time image color stretch, and contour profiling to allow the user to fully explore the data while in the immersive environment.

We have also explored digital field notebooks and their applicability on future manned space missions. We have successfully deployed test systems in the Arctic Dry Valley, and we incorporating ideas provided by Commander David Scott, Apollo 15 into our next generation of ADVISER software. This software will be complimented by a new state-of-the-art Cave immersive display system being developed at Brown University through a new $2M NSF MRI award, with high enough resolution to truly mimic an astronaut's space suit-based instrumentation within the virtual setting of a planet's surface.

In summary, we have made significant progress towards our overall goal of developing technologies that increase the science return. The tools aim to allow geoscientists to explore and investigate remote locations as though they were "virtually" there. In the third year, we plan to focus on integrating new Lunar and Martian data sets, continue to refine ADVISER as a cross-platform scalable toolset, and adapt the system to take advantage of our evolving hardware improvements.

Automated DTM Generation for HiRISE and LROC

Zachary Moratto
Stinger Ghaffarian Technologies

A demand for Digital Terrain Models (DTM) has been growing within the scientific community as hundreds of unprocessed stereo image pairs are captured by high-resolution imagers around Mars and the Moon. Traditional methods for processing these stereo pairs only incorporate a modest level of automation; many man-hours are still required to guide the software, and few images have been processed as a result. Our proposal funds the development of the Ames Stereo Pipeline, an automated stereo processing engine that can process stereo imagery with little to no human intervention. We will discuss how this pipeline has been adapted to process HiRISE and LROC imagery, and show our latest results.

In late October, the Ames Stereo Pipeline will be released freely to the public as NASA open source software. This product will interoperate with the widely used USGS Integrated Software for Imagers and Spectrometers (ISIS), thereby providing planetary scientists with a sophisticated, free application for automated terrain generation that integrates with their existing workflow.

The Cloud Goes BOOM: Data-Centric Programming for Datacenters

Joe Hellerstein
University of California, Berkeley

The Berkeley Orders Of Magnitude (BOOM) project is an effort to provide cloud developers with a programming environment to build Orders of Magnitude bigger systems in OOM less code. As a first experiement, we rebuilt Hadoop and HDFS using our earlier Overlog language. The resulting Boom Analytics stack is API-compliant and performance-competitive with Hadoop, and includes sophisticated new system features including high availability via Paxos master replication, horizontal scale-out of master nodes, and pipelined query processing for online aggregation. I'll reflect on some of the lessons we learned, particularly from our MultiPaxos implementation, which are informing the design of a new language currently being called Lincoln.

Visualization of Terascale Datasets with Impostors

Thomas Quinn
University of Washington

The interactive visualization of large astrophysical datasets is a formidable challenge. Cosmological simulations performed on current terascale facilities typically generate data sets up to 100 gigabytes in size. In order to understand the complex three-dimensional structure within these simulations, interactive visualization with rapid rotation and zooming is required, but handling this much data is well beyond the capability of even current high-end graphics workstations. The rendering can be performed on a parallel computer, even the same facility on which the simulation was done, but the user typically is using such facilities remotely. Hence both the latency and the bandwidth between the user's workstation and the parallel machine preclude the possibility of interactive rendering.

We are using techniques from the computer graphics community to overcome these latency and bandwidth hurdles. We have explored using 3D voxel grids to represent particle data. The voxels are rendered as 3D textures on the parallel server, communicated to the client where they are stored as 3D textures directly on the client's graphics card. Real-time rotation can then be performed by the graphics hardware. These techniques have been implemented within a framework that is currently being actively used in research on galaxy formation and evolution, and is sufficiently extensible that it can be used in a variety of applications relevant to NASA from planet formation simulations to catalogs of galaxies from large-scale structure surveys. Hence data from all these sources can be rendered in real-time 3D.

Large Scale On Demand Cross-Matching with Open SkyQuery

Ani R. Thakar
Johns Hopkins University

Being able to perform arbitrary cross-matches on demand between data from multi-wavelength mega surveys is the holy grail for astronomy in the digital age, and one of the driving goals of the Virtual Observatory is to make this a seamless reality. (Open) SkyQuery is a prototype we developed at JHU to federate geographically distributed astronomical archives and enable SQL queries to cross-match data between them. As a demonstration of the promise of the VO, it has been remarkably successful and is one of the most heavily used VO applications. However it currently only allows small cross-matches, due to design and performance limitations.

We are now in the process of re-engineering SkyQuery to enable it to handle arbitrary sized cross-matches from the largest surveys of today and tomorrow. This effort primarily involves a new cluster-based architecture called GrayWulf, a more efficient Bayesian cross-match algorithm, and an asynchronous workflow system to handle the movement of data. GrayWulf enables queries on large datasets to be partitioned across a cluster of SQL Server database nodes so they can be executed in parallel. The Zone algorithm is used to efficiently partition the queries. The workflow system manages the job scheduling, parallel execution, and copying and caching of remote data.

On-the-fly and Grid Analysis of Astronomical Images for the Virtual Observatory

Andrew Ptak
Johns Hopkins University

We are developing a system to combine the Web Enabled Source Identification with X-Matching (WESIX) web service, which emphasizes source detection on optical images, with the XAssist program that automates the analysis of X-ray data. XAssist is continuously processing archival X-ray data in several pipelines. We have established a workflow in which FITS images and/or (in the case of X-ray data) an X-ray field can be inputted to WESIX. If an X-ray field is requested, WESIX queries XAssist web services to see if the field is publicly available, and if so whether it has been fully processed in a pipeline. If it has been processed, then image(s) and source lists are returned. If the X-ray data is public but has not been processed, XAssist returns the X-ray image for WESIX to perform quick-look analysis (using SExtractor) and the field is added to the XAssist queue for full processing. These services will be available via web services (for non-interactive use by Virtual Observatory portals and applications) and through web applications (written in the Django web application framework). We are also adding web services for specific XAssist functionality such as determining the exposure and limiting flux for a given position on the sky and extracting spectra and images for a given region. We are also improving the queuing system in XAssist to allow for “watch lists” to be specified by users; when X-ray fields in a user’s watch list become publicly available, they will be automatically added to the queue.

Martian Tactical Geologic Mapping for Mars Surface Missions

Mark Powell
NASA Jet Propulsion Laboratory

This work focuses on new techniques for visualization and removal of error in geologic mapping for Mars surface missions. The current state of practice in Mars rover on-board localization error correction is too time-consuming and resource-intensive to be used on a frequent basis in the course of tactical operations. The current state of practice in ground-based localization error correction is too time-consuming to be integrated into the tactical activity planning schedule. New interactive visualization tools now allow for mission scientists to correct for localization error by leveraging their expert knowledge in a highly efficient user interface regime. Our visualization tool performs interactive data fusion of Mars orbital and surface imagery. With this tool, a geologist can achieve correction of Mars rover localization error by three orders of magnitude in a matter of seconds. This tool also provides a collaborative capability that allows all mission participants to take advantage of this extremely precise localization and mapping data as useful context for a wide variety of science investigations. The corrected datasets generated with this visualization have been made available to the world via the Google Mars application. In addition, we have created a powerful 3D visualization capability for tactical activity planning. This 3D visualization allows a scientist to browse a regional elevation map for a surface mission with overlays of terrain meshes built from high-resolution stereo imagery captured by rover navigation and science cameras.

Integration of Orbital, Descent and Ground Imagery for Topographic Capability Analysis in Mars Landed Missions

Rongxing (Ron) Li
Ohio State University

With support from the NASA Applied Information Systems Research (AISR) Program, the Mapping and GIS Laboratory at The Ohio State University has developed an advanced methodology for the integration of orbital, descent, and ground data for topographic capability analysis in Mars landed missions. Mars orbital (satellite), descent (spacecraft), and ground (lander/rover) data have been integrated to support lander localization and surface operations of the Mars Exploration Rover (MER) 2003 mission. The developed systematic analysis and integration capabilities of these three categories of image data are critical to the achievement of high-precision 3D mapping, rover navigation, and other science objectives.

We here present the research activities and final results for the AISR research project "Integration of Orbital, Descent and Ground Imagery for Topographic Capability Analysis in Mars Landed Missions". In this project we have developed a rigorous method of photogrammetric processing for automatic generation of topographic products from HiRISE (High Resolution Imaging Science Experiment) stereo images. Our method employs a coarse-to-fine hierarchical matching process that is able to reliably provide dense matched points for DEM generation and evenly distributed tie points for bundle adjustment (BA). A new method for automated interest point matching based on the integration of orbital- and ground-based image networks also has been developed. This method can construct an orbital image network from HiRISE stereo images based on the rigorous sensor model. A ground image network is constructed using MER telemetry data, visual odometry, and incremental BA based on photogrammetric processing of ground images. The resulting rover traverse is incorporated into the orbital image network as horizontal control. When slippage occurs and the incremental BA is not available due to lack of ground features, as in the case of Opportunity rover, landmarks identified from both orbital and ground image networks are used to adjust the rover position. We tested a new concept of automatic rock matching between orbital and ground images that can facilitate a new way of orbital image guided/verified rover localization. Test results show that the simplified, efficient methodology that has been developed can effectively extract rocks and adjust rover positions at a level of precision comparable to that achieved by the more labor intensive incremental BA currently used for the MER mission. This method opens up a new possibility for autonomous rover localization performed onboard unmanned vehicles for future space exploration.

GPU-based Tools for Computational Astrophysics: N-Body Integrators for Dynamical Systems

Mario Juric
Institute for Advanced Studies

Graphics Processing Units (GPUs), have recently emerged as powerful and cost-effective accelerators of arithmetically intensive codes, when coupled with appropriately parallelized algorithms. To leverage this resource in computational astrophysics, we are developing efficient GPU-based algorithms for the parallel integration of systems of ordinary differential equations. The developed tools should be particularly valuable for any problem requiring the exploration of high-dimensional spaces for model parameters, global search, model fitting, estimating parameter uncertainties, and evaluating the robustness of conclusions to uncertainties.

Our scientific goal is to initially apply them to the difficult problem of characterization of the dynamics of exoplanetary systems. In particular, we plan to develop a highly parallel n-body integrator optimized for studying planetary systems. Based on estimates and initial experiments, we expect that our GPU-based kernel for n-body integration will achieve nearly two orders of magnitude greater performance than a comparable single-threaded CPU-based code.

Providing these tools to the scientific community will significantly expand the research capabilities in planetary dynamics, as well as computational astrophysics in general.

Controlling Sensitive Trajectories, Mission Extension, and Material Transfer Between Planetary Systems

Edward Belbruno
Princeton University

Recent results have shed light on the nature of the weak stability boundary region about the Moon and other bodies. This has been an elusive problem for many years since this region was discovered in 1986. Now, we can begin to understand this region with a newly refined algorithmic definition showing it has a very complicated fractal structure. This is illustrated with computer visualizations. These results imply that the WSB is closely associated with the manifolds (tubes) associated with the Lyapunov orbits about the Lagrange points. These results are very surprising. They also show how to estimate a hyperbolic tangle without the use of invariant manifolds for the first time. Recent results of another problem are also described on showing how the probability of transfer of material between planetary systems can be enhanced when utilizing low-energy transfers associated to the WSB regions about the associated stars.

Hadoop: Distributed Data Processing

Amr Awadallah
Cloudera

Hadoop is an open-source distributed data processing system designed to handle petabyes of structured and unstructured data at extremely low cost—essentially the cost of the clustered, commodity hardware upon which Hadoop runs. Hadoop is Apache's implementation of the MapReduce/GFS frameworks popularized by Google. Yahoo invested significant resources in the Hadoop project to handle many large data processing tasks, including analyzing and monetizing trends in atomic-level clickstream data. In this talk we will take you to the future of data processing, in which there are no limits to the amount of data an organization can store and analyze cost effectively.

Development of an Adaptive Non-Ideal MHD Simulation Tool for Multiple Space Science Applications

Gabor Toth
University of Michigan

Over the last 16 years our group at the University of Michigan has been developing a general use global MHD code, BATS-R-US, and the Space Weather Modeling Framework (SWMF) that couples domain models extending from the Sun to planetary upper atmospheres and ionospheres. BATS-R-US and the SWMF have been extensively used to simulate a broad range of space science phenomena. Still there are many unmet challenges. There is a need to go beyond ideal MHD and to improve the efficiency of the numerical schemes.

In the first year of this AISRP we have:

- finished and published the Hall MHD scheme
- added an electron equation to the MHD equations
- added a multi-ion MHD equation module
- tested the empirical resistivity model by M. Kuznetsova

In the second year we have:

- added the electron pressure gradient term to Hall MHD
- finished the multi-ion MHD implementation and published a paper
- implemented a prototype for the time-accurate local time stepping algorithm
- started work on the non-isotropic pressure
- implemented a new algorithm to reduce the numerical diffusion

I will briefly describe the new developments in my talk.

Parallel-Processing Astrophysical Image-Analysis Tools

Kenneth Mighell
National Optical Astronomy Observatory

The AISR-funded enabling image-processing technology of the Principal Investigator's MATPHOT algorithm, which does precise and accurate stellar photometry and astrometry with discrete (sampled) Point Spread Functions, has been improved and extended for use with space-based near-infrared cameras with lossy detectors. The PI has demonstrated that the precision of astrometry and aperture photometry of bright isolated stars observed with Channel 1 (3.6 micron) of the Infrared Array Camera (IRAC) instrument of the Spitzer Space Telescope can be calibrated to produce results that are comparable to that produced by Point-Spread-Function-fitting procedures. An optical model has been developed which explains the systematic centroid error of IRAC Ch1 point-source observations measured with standard intensity-weighted centroid algorithms. This methodology can be applied to investigate possible systematic centroid errors that may be present in the newly installed WFC3 instrument of the Hubble Space Telescope and the NIRCam instrument which is being built for the James Webb Space Telescope.

EPISODE: Software for Trajectory Generation

Jeff Jewell
NASA Jet Propulsion Laboratory

Trajectory generation for the three and higher-body problem is a notoriously difficult constrained optimization problem. The inherent difficulty in finding optimal solutions lies in the presence of many local minima of the objective function (i.e., fuel or time), and consequently, convergence to optimal solutions by deterministic algorithms require very good initial guess solutions. The technical innovation explored in this project includes a probabilistic algorithm for trajectory generation (implemented computationally with the code EPISODE), which provably converges to globally optimal solutions. Two key components to this approach are: 1) the ability to generate a large class of possible trajectories in a semi-automatic fashion utilizing the "interplanetary superhighway" (IPS) (a network of exact solutions of the three and higher-body dynamics which asymptotically wind on and off periodic orbits), and 2) the ability to evaluate a trial trajectory's "probability" of leading to an optimal solution with continued computing. This computational approach to the trajectory design problem is potentially enabling of much more complicated missions involving low-thrust (ion) propulsion, which visit a large number of science "targets of opportunity."

How Well Do You Know That? Uncertainty Analysis in Earth Remote Sensing

Robin Morris
RIACS

We present progress on our project to accurately quantify the uncertainty on biospherical data products produced from remote sensing data. We concentrate on the production of LAI estimates by inverting a canopy radiative transfer model. We present results of performing a fully Bayesian global sensitivity analysis of the coupled Leaf Canopy Model (LCM), including estimates of the uncertainties of the main effects, and, in particular, new methodology for the estimation of the full posterior distributions of the sensitivity indices. We also study the impact of informative prior distributions over the inputs to the LCM derived from the LOPEX database, showing how the information in the database increases the sensitivity of the model to the inputs of interest, which should allow for more accurate estimation. We conclude with plans for future work.

Astronomy in the Cloud

Jeffrey P. Gardner
University of Washington

What limits astronomy today is no longer obtaining or storing data, it is the question of how we interact with and analyze Petabyte-scale data repositories. Serial access to data sets via SQL queries will not scale to the size of the science questions we wish to address. We are reaching a stage where the data are much richer than the analyses we apply to them. Astronomy is, therefore, in need of a new model for knowledge discovery with massive data sets; one where the analysis moves to the data rather than from the database to the user, where the application scales elastically to the size of the problem in hand, and where the figure of merit is not just the efficiency of the algorithm but the time between posing the question of the data and receiving the answer.

We are exploring how the emerging map-reduce paradigm for data-intensive computing might revolutionize the way that the sciences approach access to and analysis of a new generation of massive data streams. We focus on a series of astrophysical challenges, using Hadoop to develop an image analysis framework for the spectral and temporal analysis of massive astronomical image data sets. We will address a number of computationally challenging problems: the spatial distribution of star formation within resolved galaxies and the search for moving objects within the outer solar system. In terms of a broader impact, we expect our techniques will extend beyond astrophysics to other data intensive sciences. If successful, this will add a new scientific paradigm to a broad range of applications that can scale from a few multicore compute nodes up to national high-performance computing facilities.

Directed Exploration of Complex Systems

Michael Burl
NASA Jet Propulsion Laboratory

Physics-based simulation codes are widely used in science and engineering to model complex systems that would be infeasible to study otherwise. Such codes provide the highest-fidelity representation of system behavior, but are often so slow to run that insight into the system is limited. As one example, we are using an SPH/N-Body asteroid collision simulation to understand the conditions that could produce a Karin-like asteroid family. Conducting an exhaustive sweep through the input parameter space is too expensive; instead, we consider a *directed exploration* strategy, in which the next simulation trials are cleverly chosen at each step based on the results of previous trials. Essentially, we use the simulator as an oracle to build up a body of labeled training data (input-output pairs). A supervised learning algorithm ingests the available training data and forms a simplified predictive model of the system. Active learning is then used to identify which simulation trials would be most valuable to run next. Processing the new trials through the simulator produces new training data. The process is then repeated until some stopping criteria is met. Choosing good trials at each step is critical due to the high time cost associated with the simulation code.

Demonstration of Safe Human/Robot Coordination on the Athlete Lunar Rover

Brian C. Williams
Massachusetts Institute of Technology

In future human exploration missions, robotic systems will work together with human astronauts to complement their physical limitations. In some cases these robots will need to work safely while in close proximity to their human counterparts. One such example is the Athlete Lunar Rover, which augments astronaut teams with a highly dexterous heavy lifting capability. Our objective is to develop robust, cooperative execution capabilities that allow human/robot teams to work as effectively and safely together as human teams.

Effective human teams are highly adaptive. For example, an effective scrub nurse works hand to hand with a surgeon, while assessing and anticipating the surgeon's needs, responding quickly to changing circumstances, and responding quickly to the surgeon's cues and requests. A teammate must also be physically deft with his or her environment; for example, quickly recovering after tripping or bumping into objects. In this talk we explore the problem of task coordination and execution between a robot and its human teammate under time pressure. We address two problems: First, how can a robot make choices about its course of action, without overly constraining its teammate's options, and while permitting its teammate to make choices on the fly? Second, how can a human-like robot execute its tasks successfully, by meeting all desired deadlines, while achieving compliance in motion?

We frame this problem as a form of multi-agent temporal plan execution. We build upon prior work on dynamic execution of temporal plans, in order to make decisions dynamically, while adapting to uncertain events that are out of the control of the robot. To support teamwork, we extend dynamic execution to adapt to uncertainty regarding the tasks that its teammate will choose. To support motion compliance we extend dispatchable execution to continuous dynamic control of the robot's limbs. We validate this work in simulation and in hardware on several human-like robot platforms, including the Vecna Bear robot simulation, a multi-manipulator coordination testbed at MIT CSAIL, and the Athlete lunar rover, within the Mars Yard at JPL.

This work is in collaboration with Andreas Hofmann, Julie Shah, Patrick Conrad, and Shannon Dong at MIT CSAIL, and with David Mittman, Mitch Ingham and Vandi Verma at Caltech JPL.

CIDU Posters

Many of these posters can be found at https://dashlink.arc.nasa.gov

A Text Classification Approach Using Semi Supervised Subspace Clustering
Mohammad Salim Ahmed, University of Texas at Dallas

Active Label Generation and Stream Mining Approaches To ASRS Anomaly Reports
Clay Woolam, Mohammed Masud, Latifur Khan, University of Texas at Dallas

Aircraft Trajectory Clustering
Maxime Gariel, Georgia Tech

Applying Textual Entailment to Cause Identification from Aviation Safety Reports
Muhammad Abedin, University of Texas at Dallas

Automatic Coronal Loop Detection from SOHO/EIT Images
Nurcan Durak, University of Louisville

Classification Models for Data Labeling
Clay Woolam/Mohammed Masud/Latifur Khan, University of Texas at Dallas

Data Mining of Air Traffic Track Data for NGATS
Johann Schumann, RIACS/USRA

Feature Transformation Using Wavelets: Application to Despiking for Diagnostic Systems
Ehsan Sheybani, Virginia State University

Finding Chemical Compounds on Enceladus with Correlation Absorption Signatures
Rachel Mastrapa/Bryan Matthews, NASA Ames Research Center

From Mining for the Unusual in Complex Systems to Knowledge Discovery in High-Priority Applications
Auroop R. Ganguly, Oak Ridge National Laboratory

From Structural Health Monitoring to Structural Health Management: A Road Map
Cecilia Larrosa, Stanford University

Identifying Cause Factors in ASRP Safety Reports
Isaac Persing, University of Texas at Dallas

MiTexCluster: Micro-Text Cluster Cube for Online Summarization of Text Cells
Duo Zhang, University of Illinois at Urbana-Champaign

Multidimensional Text Databases
Yintao Yu, University of Illinois at Urbana-Champaign

Online Data Mining Framework for the Electric Power Grid
Olufemi A. Omitaomu, Oak Ridge National Laboratory

Parameter Estimation of ERIS model using Monte Carlo Error propagation Methods
Rajkumar Thirumalainambi, PSGS

Porting Gravitational Wave Signal Extraction to Parallel Virtual Machine (PVM)
Rajkumar Thirumalainambi, PSGS

Scalable Pathfinder Algorithms to Support Text Mining Tasks
Samson Hauguel, University of Illinois at Urbana-Champaign

Searching Top-k Cells in Text Cube
Bolin Ding, University of Illinois at Urbana-Champaign

Solving a Prisoners' Dilemma in Distributed Anomaly Detection
Nisheeth Srivastava, University of Minnesota/Aleksandar Lazarevic, UTRC

Virtual Sensor Web Infrastructure for Collaborative Sensing (VSICS)
Prasanta Bose, Lockheed Martin Space Systems

AISRP Posters

Many of these posters can be found at https://dashlink.arc.nasa.gov

A Novel Scheme for the Compression and Classification of Hyperspectral Images
Tamal Bose Bei Xie/Erzsebet Merenyi, Virginia Tech

Conjoined Twins: A Smart Neural Architecture for Adaptive Inference of Multiple Latent Parameters from Complex High-Dimensional Data
Erzsbet Mernyi/Lili Zhang, Rice University

Discriminating Patterns for Surveying and Mapping Geospatial Relationships
Wei Ding, University of Massachusetts

Flexible Probabilistic Modeling of Astrophysical Dynamic Spectra
Thomas Loredo, Cornell University

Global Image and Elevation Datasets for Computer Graphics using the Gnomonic Projection
Andrew Loomis, Brown University

Integration of Orbital, Descent and Ground Imagery for Topographic Capability Analysis in Mars Landed Missions
Rongxing Li, The Ohio State University

Multi-modal Image Registration for Titan Balloons
Larry Matthies/Adnan Ansar, NASA Jet Propulsion Laboratory

Multi-Objective Multi-Participant Scheduling for Space Science Missions
Mark Johnston/Mark Giuliano, Space Telescope Science Institute

Robust Grid Computing using Peer-to-Peer Services
Alan Sussman, University of Maryland

Solar Loop Mining to Support Studies of the Coronal Heating Problem
Olfa Nasraoui, University of Louisville

Virtual Planetary Analysis Environment for Remote Science
Leslie Keely, NASA Ames Research Center